MISCELLANEA.

I. On Spurious Values of Intra-class Correlation Coefficients arising from Disorderly Differentiation within the Classes.

By J. ARTHUR HARRIS, Ph.D. Carnegie Institution of Washington, U.S.A.

When the constants of the x and y characters of the population in r_{xy} are quite indistinguishable symmetrical tables* may be used, but not otherwise.

Primarily and for the most part, however, the use of symmetrical tables has been restricted to cases in which the degree of interdependence between the measures of all possible pairs† drawn from a considerable series of associated individuals—in short to intra-class correlations‡—is sought.

The dangers of spurious correlation due to the artificial symmetry of the surface is then much greater§. Pearson‖ long ago pointed out that when intra-class differentiation exists, for example, because of age in the case of characters determined upon the members of a fraternity, or of position on the axis in the case of serial organs, the values of r may be to some extent spurious.

In the cases considered by Pearson differentiation is an orderly phenomenon, i.e. the magnitudes under consideration increase or decrease with age, position on the axis, or some other extrinsic characteristic with such regularity that the relationship can be expressed by an equation which may be used in correcting the raw values of r.

In other cases, the problem is not so simple. Differentiation within the class may exist, but it may be difficult or impossible to arrange the individual measurements by any character outside of themselves to obtain the constants necessary for determining the true correlations from the spurious values deduced from the tables.

Illustration I. The correlation between yields of wheat in variety testing.

In variety testing, the experimenter seeks (or should seek), among other things, to determine the correlation between yields of varieties in different years. If this correlation be 0 (and regression be linear) it is clear that the yield of a variety in one year furnishes no basis for prediction

* R. Pearl, *Biometrika*, Vol. v. pp. 249—297, 1907; H. S. Jennings, *Journ. Exp. Zool.* Vol. xi. pp. 1—134, 1911 ; J. Arthur Harris, *Biometrika*, Vol. vii. pp. 325—328, 1910.

† K. Pearson and others, *Phil. Trans.*, A, Vol. cxcvii. pp. 285—379, 1901; K. Pearson and A. Barrington, *Eugenics Laboratory Memoirs*, No. V, 1909.

‡ *Biometrika*, Vol. ix. pp. 446—472, 1913.

§ With only one pair of measures the probability of spurious correlation is, in cautious work, very slight, for the possibility of differentiation can be easily tested by the critical comparison of the physical constants.

‖ Pearson, K., "On Homotyposis in Homologous but Differentiated Organs." *Roy. Soc. Proc.* Vol. lxxi. pp. 288—313, 1903.

concerning its yield in any subsequent year. If, on the other hand, the correlation be high, prediction from a few years' test may be made with great probability of certainty.

Given a measure of the "performance" of a series of varieties during a number of years it would at first seem quite allowable to form symmetrical tables or to use the intra-class formulae of a former paper* to determine the intra-varietal correlation, and to regard this as a satisfactory measure of the differentiation of the varieties and of the average prediction value of a year's test. Such is, however, not the case, for while there may be no orderly change in yield throughout the period under consideration, the individual years differ greatly in their average yield for all the varieties. The influence of this "disorderly differentiation" upon r is admirably shown by A. D. Hall's† table of the yield in bushels of wheat in the Rothamsted experiments.

Let b = yield in bushels per acre of any one of m varieties in any one of n years, y_1, y_2 be the "first" and the "second" years of a symmetrical intra-varietal correlation surface, v_1, v_2 be the "first" and "second" varieties of a symmetrical intra-annual correlation surface. Then $r_{b_{y_1} b_{y_2}}$ will be a (spurious) measure of the (persistent) differentiation of varieties, $r_{b_{v_1} b_{v_2}}$, a (spurious) measure of the differentiation (in the yield of all the varieties) of years. Applying formulae (v)—(ix) of *Biometrika*, Vol. IX. p. 450, to these data, I find

$$S[n(n-1)] = 2128,$$

$$S[(n-1)\Sigma(b')] = 83122{\cdot}5, \quad S[(n-1)\Sigma(b'^2)] = 3483626{\cdot}4,$$

$$S[\Sigma(b')]^2 = 3610204{\cdot}57, \quad S[\Sigma(b'^2)] = 370820{\cdot}13,$$

$$\bar{b} = 39{\cdot}0613, \quad \sigma_b^2 = 111{\cdot}257328,$$

$$r_{b_{y_1} b_{y_2}} = -{\cdot}032.$$

The result is obviously spurious, for mere inspection of the entries in the table shows that some varieties regularly give heavier yields than others. The source of the spurious value is to be seen in the fact that an intra-class coefficient has been calculated from a symmetrical surface formed from classes (varieties) represented by a series of yields *differentiated by annual variations in the growing conditions*. By correcting for this source of differentiation by expressing each yield as a deviation from the mean yield of all the varieties for the particular year, i.e. $b'' = b - \bar{b}_y$, where the bar denotes a mean and the subscript y that it is for all the yields of a year, I have found‡

$$r_{b''_{y_1} b''_{y_2}} = {\cdot}266.$$

Measuring the differentiation of years in terms of intra-annual correlation (intra-class correlation in which each class is defined by the year and its individuals are the yields of the different varieties grown), I find from Hall's table

$$S[m(m-1)] = 4440,$$

$$S[(m-1)\Sigma(b')] = 174129{\cdot}2, \quad S[(m-1)\Sigma(b'^2)] = 7317531{\cdot}92,$$

$$S[\Sigma(b')]^2 = 7586436{\cdot}21, \quad S[\Sigma(b'^2)] = 370820{\cdot}13,$$

$$\bar{b} = 39{\cdot}2183, \quad \sigma_b^2 = 110{\cdot}017719,$$

$$r_{b_{v_1} b_{v_2}} = {\cdot}791.$$

Since the varieties have been shown to be differentiated, this result must also be spurious. Let $b'' = b - \bar{b}_v$ where the v indicates that the mean denoted by the bar is for the yield of the

* *Biometrika*, Vol. IX. pp. 446—472, 1913.

† Hall, A. D., *The Book of the Rothamsted Experiments*, p. 66, 1905.

‡ *Science*, N. S. Vol. XXXVI. pp. 318—320, 1912. Probably a better method of dealing with such cases will sometime be found. So far I have not succeeded.

TABLE I.

Yield of Varieties of Wheat in Different Years.

Variety	1871	1872	1873	1874	1875	1876	1877	1878	1879	1880	1881	Lots	$\Sigma(y)$	$\Sigma(y^2)$
Rivet (Red)	—	—	48·1	67·0	48·4	42·5	49·6	66·1	16·0	22·4	52·2	9	412·3	21263·39
White Chaff (Red)	—	—	40·6	55·1	40·2	49·5	48·4	59·0	22·8	28·1	54·5	9	398·2	18853·92
Club Wheat (Red)	36·0	45·8	47·5	59·6	46·6	47·6	49·5	61·0	23·5	16·4	43·4	11	476·9	22515·39
Golden Drop (Red), Hallett's	39·5	49·8	44·2	51·8	38·1	48·4	49·5	52·8	21·0	18·9	50·8	11	464·8	21088·28
Bole's Prolific (Red)	33·6	42·8	45·2	48·1	43·8	41·4	44·8	52·8	31·0	24·5	46·5	11	454·5	19468·23
Hardcastle (White)	—	46·5	42·0	49·6	33·9	44·0	42·1	54·0	21·5	24·4	45·6	10	403·6	17297·00
Red Rostock	37·0	—	46·3	53·8	37·4	40·0	46·4	57·0	8·5	28·4	45·8	10	400·6	17784·30
Red Langham	30·8	43·8	34·1	53·1	34·9	42·5	42·9	50·8	25·8	28·6	48·5	11	435·8	18130·66
Bristol Red	29·4	44·4	39·5	53·4	31·6	42·4	44·1	52·1	21·6	30·6	46·2	11	435·3	18240·43
Red Wonder	31·2	43·8	37·1	55·1	33·2	44·2	41·6	52·1	22·0	28·2	45·9	11	434·4	18191·20
Red Chaff (White)	32·8	37·0	35·3	48·8	34·3	43·8	41·0	—	—	—	—	7	273·0	10848·30
Browick (Red)	35·3	40·5	38·5	51·1	38·5	39·1	40·9	49·5	24·0	19·6	47·3	11	424·3	17311·37
Casey's White	29·9	42·1	37·5	52·1	39·0	45·5	43·0	47·8	15·4	24·1	42·9	11	419·3	17170·55
Red Nursery	34·1	45·3	27·1	41·1	39·0	37·5	40·6	47·8	30·9	27·5	46·0	11	416·9	16326·03
Wooly Ear (White)	31·2	42·8	37·0	51·3	36·1	46·6	37·5	48·3	20·0	21·0	44·1	11	415·9	16805·69
Burwell (Old Red Lammas)	31·1	41·3	35·1	47·3	38·5	38·4	39·0	46·3	27·0	27·0	44·8	11	415·8	16228·74
Golden Rough Chaff (Red)...	33·0	39·3	38·5	52·1	38·8	38·4	36·4	46·8	14·4	31·3	41·6	11	410·6	16242·96
Chubb Wheat (Red) ...	28·4	40·0	35·8	50·5	38·3	40·3	41·5	55·1	20·8	14·9	—	10	365·6	14742·34
Original Red (Hallett's) ...	30·0	35·3	36·4	43·6	26·0	40·1	44·4	—	—	—	—	7	255·8	9627·38
Victoria White (Hallett's) ...	33·8	45·3	38·3	44·3	33·8	41·1	42·6	43·9	14·9	15·8	44·0	11	397·8	15605·18
White Chiddam	26·9	38·8	31·8	42·0	32·4	37·5	37·6	49·8	11·9	27·4	47·1	11	383·2	14464·88
Hunter's White (Hallett's) ...	26·9	39·8	38·0	45·4	26·4	43·5	40·0	42·3	17·4	22·8	—	10	342·5	12613·91
Number of Lots	19	19	22	22	22	22	22	20	20	20	18	226	—	370820·13
Total Yields, $\Sigma(y)$	610·9	804·4	853·9	1116·2	809·2	934·3	943·4	1035·3	410·4	481·9	837·2	—	8837·1	—

variety for all the years it was grown. Correcting for the influence of the differentiation of varieties in this way I have found*

$$r_{b''_{p_1}b''_{p_2}} = \cdot 837.$$

Thus season is a far more important factor than variety in determining an individual yield.

Illustration II. Influence of Personal Equation upon the Correlation between the Grades assigned to the Same Paper by a Series of Instructors.

Stripped of the verbiage in which it has been clothed in discussions among pedagogues, one of the chief problems concerning the reliability of the grades assigned in examinations resolves itself unto the statistical question: What is the correlation between the grades assigned to the same paper by different instructors?

Let g be the grade assigned to any one of m papers by any one of n instructors, let i_1, i_2 be the "first" and "second" instructor (of a symmetrical intra-class table) passing judgment upon a paper, p_1, p_2 the "first" and "second" paper graded by the same instructor. Then from Table I of D. Starch† I deduce, by the intra-class formulae (v)—(lx) of *Biometrika*, Vol. IX. p. 450,

$$r_{g_{i_1}g_{i_2}} = \cdot 659, \quad r_{g_{p_1}g_{p_2}} = \cdot 071.$$

By using the deviation method as illustrated above, I have found

$$r_{g''_{i_1}g''_{i_2}} = \cdot 732, \quad r_{g''_{p_1}g''_{p_2}} = \cdot 386.$$

TABLE II.

Grades of Papers Assigned by Various Instructors.

Papers.	Instructors.										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	$\Sigma(g)$
1	85	86	88	85	75	80	88	87	85	87	846
2	77	80	87	80	62	82	82	87	85	87	809
3	74	78	78	75	69	84	91	83	79	80	791
4	65	65	62	20	26	60	55	68	55	50	526
5	68	82	78	82	64	88	85	86	78	80	791
6	94	87	93	87	83	77	89	88	88	89	875
7	88	90	95	87	79	85	96	91	87	89	887
8	80	84	73	79	72	83	85	91	77	76	800
9	70	70	68	50	44	65	75	81	79	79	681
10	93	92	85	92	81	83	92	89	84	85	876
$\Sigma(g)$	794	814	807	737	655	787	838	851	797	802	7882

Both of these results, in which an attempt was made to correct for the personal equation of the instructors in determining the correlation between the estimates of different instructors on the same paper, or to correct for the differences in merit of the papers in testing the individuality of the instructors, are higher than the raw values given above, which are clearly spurious. Similar results‡ are obtained from Jacoby's astronomical grades§.

* *Science, loc. cit.*

† *Science*, N. S. Vol. XXXVIII. p. 630, 1913.

‡ Personally, I can attach little pedagogical significance to series as short as those of either Starch or Jacoby. They serve here as illustrations of method merely because I know of no more extensive series.

§ *Science*, N. S. Vol. XXXI. p. 819, 1910.

The essentials of this note may be summarized as follows:

In using Intra-class coefficients care must be taken to guard against spurious values arising through differentiation among the individuals of the class.

Besides the orderly differentiation (due to age of individuals, position of organs on axis, etc.) for which Pearson has determined corrective formulae in terms of correlation coefficients, a disorderly differentiation for which such corrective formulae have not as yet been found sometimes obtains. Illustrations of such cases are here given.

Probably the empirical methods used here in correcting for this disorderly differentiation should be replaced by formulae with a sounder theoretical foundation. This I have not as yet been able to do.

The purpose of this note will have been served if it directs attention to a source of danger which may sometimes be encountered in the use of serviceable formulae, and indicates a method by which in the absence of more perfect methods practical results may be secured.

COLD SPRING HARBOR, N.Y.
February 3, 1914.

II. On an Extension of the Method of Correlation by Grades or Ranks.

BY KARL PEARSON, F.R.S.

In a memoir published in 1907* I have shown how, on the hypothesis of normal distribution, the true correlation of variates r may be ascertained from the correlation ρ of grades. If g_1 and g_2 be the two grades, ν_1 and ν_2 the corresponding ranks, x and y the corresponding variates with means $\bar{x}$ and $\bar{y}$, and standard-deviations σ_1 and σ_2, while

$$z=\frac{N}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}-\frac{2rxy}{\sigma_1\sigma_2}+\frac{y^2}{\sigma_2^2}\right)}$$

is the normal frequency surface of the variates, then

$$\bar{g}_1=\tfrac{1}{2}N=\bar{g}_2,\quad \sigma_{g_1}^2=\sigma_{g_2}^2=\tfrac{1}{12}N,$$

$$g_1-\bar{g}_1=i_1=\frac{N}{\sqrt{2\pi}\sigma_1}\int_0^x e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}dx,$$

$$g_2-\bar{g}_2=i_2=\frac{N}{\sqrt{2\pi}\sigma_2}\int_0^y e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}}dy,$$

$$\nu_1=\nu_2=g+\tfrac{1}{2},\quad \bar{\nu}_1=\bar{\nu}_2=\tfrac{1}{2}(N+1),$$

$$\sigma^2_{\nu_1}=\sigma^2_{\nu_2}=\tfrac{1}{12}(N^2-1).$$

Further I showed in the memoir just cited that

$$r=2\sin\left(\frac{\pi}{6}\rho\right)$$

* "On Further Methods of Determining Correlation," *Drapers' Company Research Memoirs* (Dulau and Co.), pp. 11, 12.